Metagenomics

Metagenomics (also referred to as environmental and community genomics) is the genomic analysis of microorganisms by direct extraction and cloning of DNA from an assemblage of microorganisms.
(Handelsman, 2004)
"Metagenomics" describes the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample
(Riesenfeld, 2004)

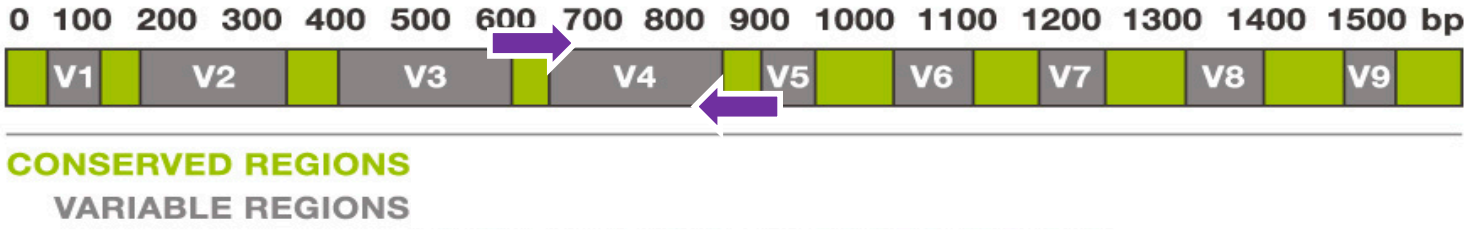
1) cultivation-independent

2) environmental and community genomics

16S rRNA gene amplicon sequencing VS. shotgun sequencing

	16S rRNA gene (or ITS) amplicon sequencing metataxonomy	Shotgun sequencing Whole genome sequencing metagenomic
Pros	• High sensitivity • Low cost	• Whole genome includes functional information • All organisms
Cons	• No functional information • Selection of Bacteria / Archaea (16S) or Fungi (ITS)	• Needs higher depth • Can be waste of information if you only use few functions

16S rRNA gene→microbiota analysis



6 *Prevotella*

4 *Clostridium*

2 *Succinivibrio*

4 *Roseburia*

2 *Turicibacter*

1 *Escherichia*

1 *Mitsuokella*

Bins!
- OTU

Things to consider when you design metagenome research experiment

- Platform?
- How many samples?
- Sequencing Depth?
- Sequencing Cost?
- Insert size? (250 ~ 500 bp?)
- Sequence length? (100 bp, 150 bp?)
- Single end, Paired end?

Sequencing cost and experimental design

HiSeq 3000: 5 Billion paired end reads
HiSeq 4000: 10 Billion paired end reads

Let us say, 8 billion per flow cell

1 billion seqs per lane
1,000,000,000 seqs

500 Milion seq per R1, R2

If you have 50 samples,
You will have 10 Million reads
Per sample (both R1 and R2)

If you have 10 samples,
You will have 50 M read per sample

Example cost:
$200 for library prep per sample
$3000 per one lane

50 x $200 = $10,000
+ $3,000 = $13,000
$260 per sample

10 x $200 = $2,000
+$3,000 =$5,000
$500 per sample

How to annotate?
Assembly or raw read?

1) Assembly based annotation (This tutorial)

Raw read

contigs

Pro: specific and longer Reference
Cons: Not all sequences can be assembled

2) Raw read based annotation (MG-RAST)

Raw read

Reference gene

BLAST to reference

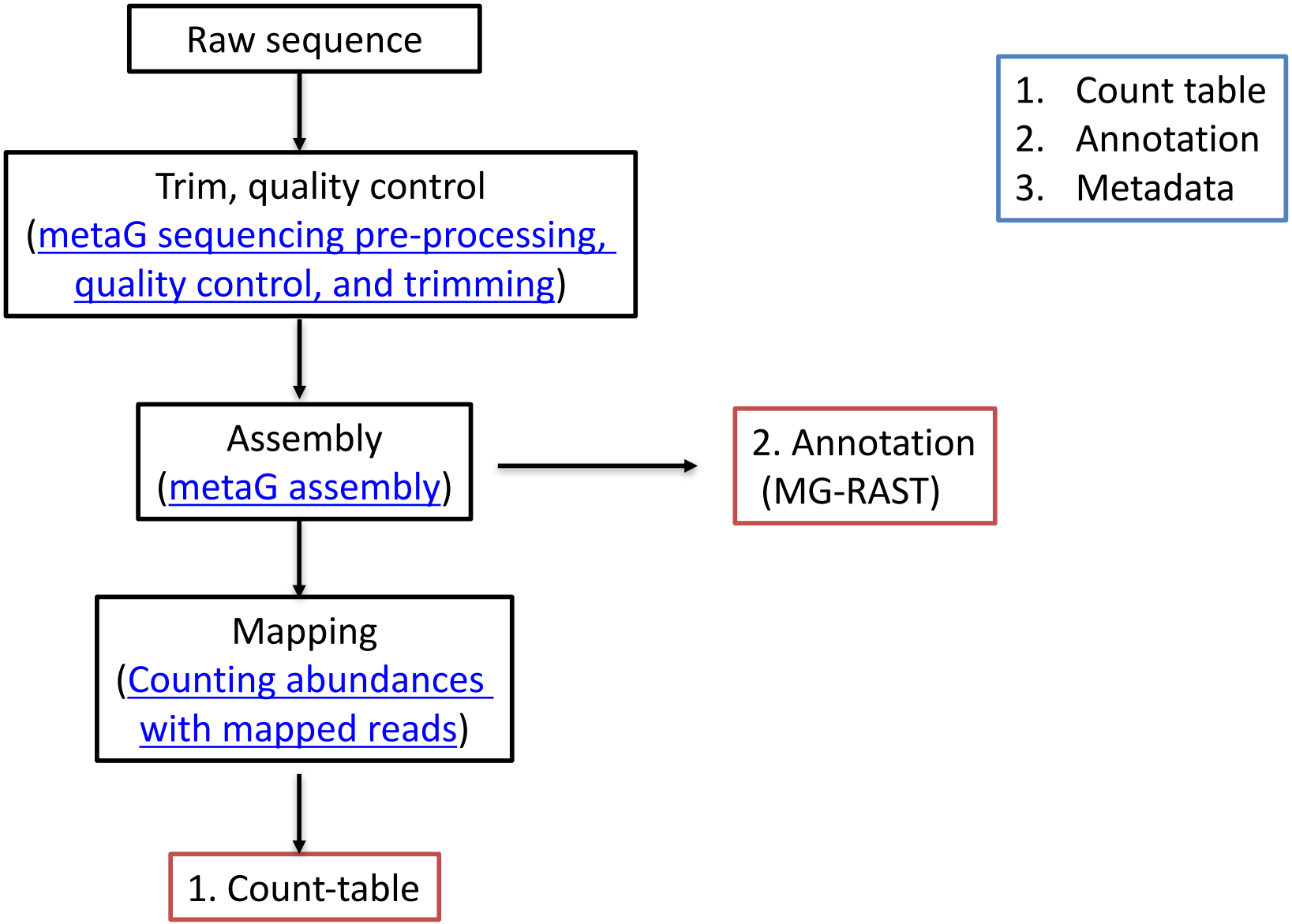
Function x
Function y
Function z
…

Pro: More sequences can be annotated
Cons: Short matching may lead Mis-annotation

Three pieces of information you need

	1. Count table	2. Annotation	3. Sample data (meta data)
16S rRNA gene Amplicon sequencing	OTU table XXX.Shared (Mothur)	Taxonomy assignment XXX.cons.taxonomy	Sample ID with other parameters
Shotgun whole genome seqeuncing	Count table	Annotation MG-RAST BLAST	Sample ID with other parameters

Overall process analyze metagenome data



The tutorial is designed for metagenomics

- But it could be used for…
  - Assembly tutorial can be used for Whole genome assembly (pure culture)
  - Mapping and counting tutorial can be used for (meta) Transcriptomics (differential gene expression)

Instruments (Platform)

What is the Flow Cell

Flow Cell for HiSeq

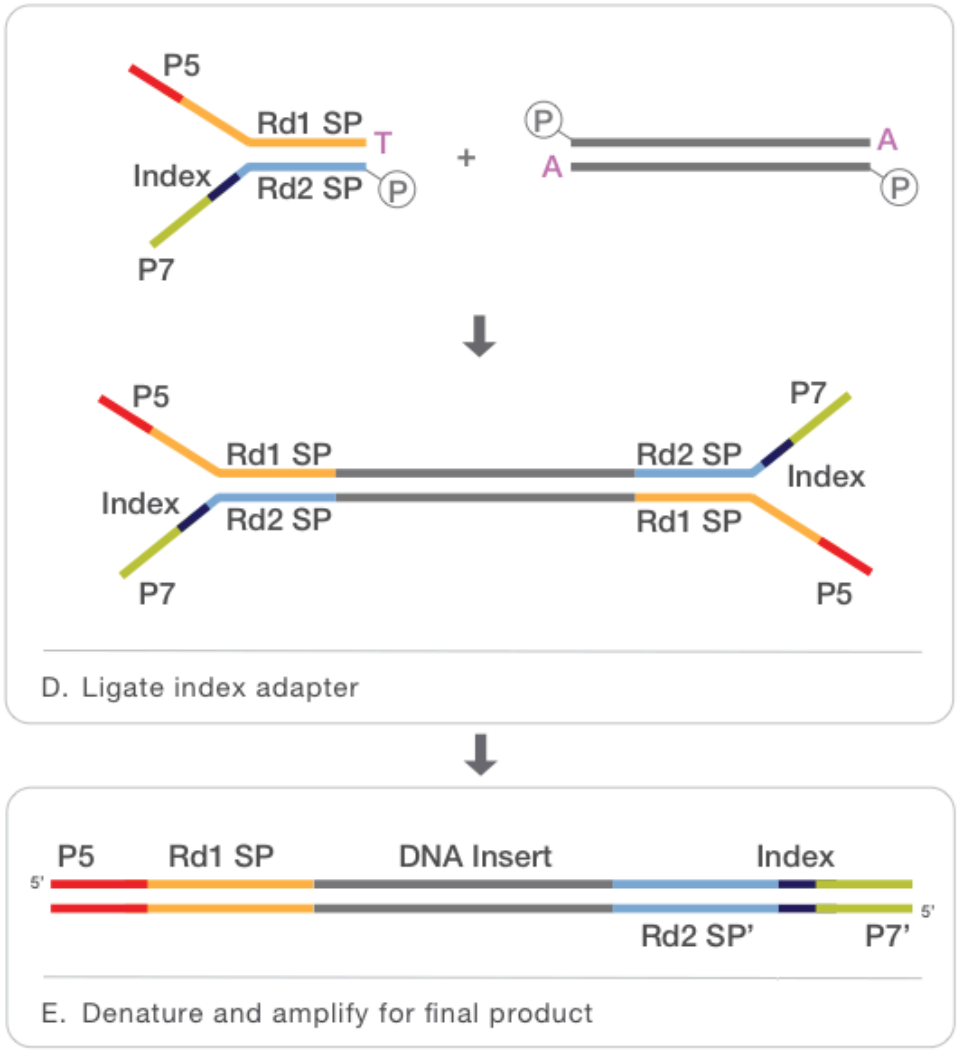
8 Lane

2 Lane

Flow Cell for MiSeq

HiSeq
https://www.youtube.com/watch?v=KQrjGiqEAq8
MiSeq
https://www.youtube.com/watch?v=tuD-ST5B3QA
Flow cell
https://www.youtube.com/watch?v=pfZp5Vgsbw0

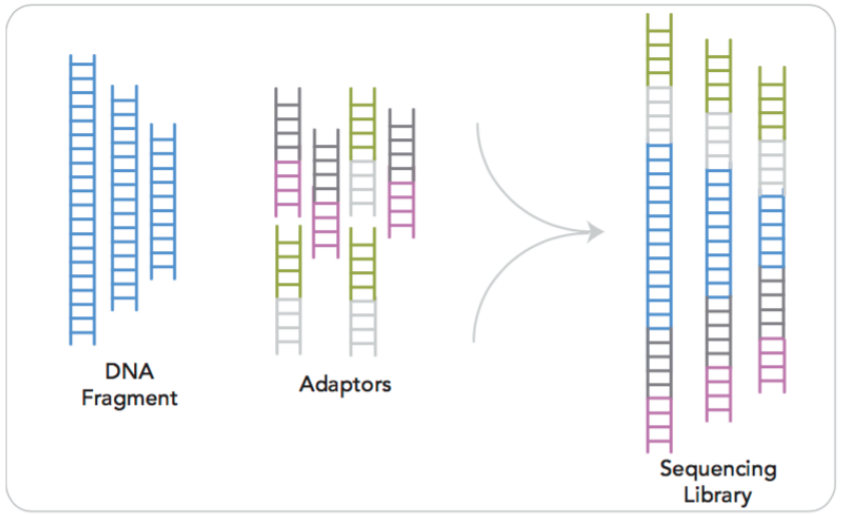
Library Preparation: Shotgun



- Adapter ligation
  - T-overhangs
  - Forked structure controls orientation
- Library amplification
  - Few cycles
  - Enrich for correctly-adapted fragments
  - Required to complete adapter structure in some protocols
- Size selection
  - Gel excision, AMPure beads
  - Limit insert size as needed, remove artifacts

What is the adapter?

Figure 1 Sequencing Library after Paired-End Sample Preparation



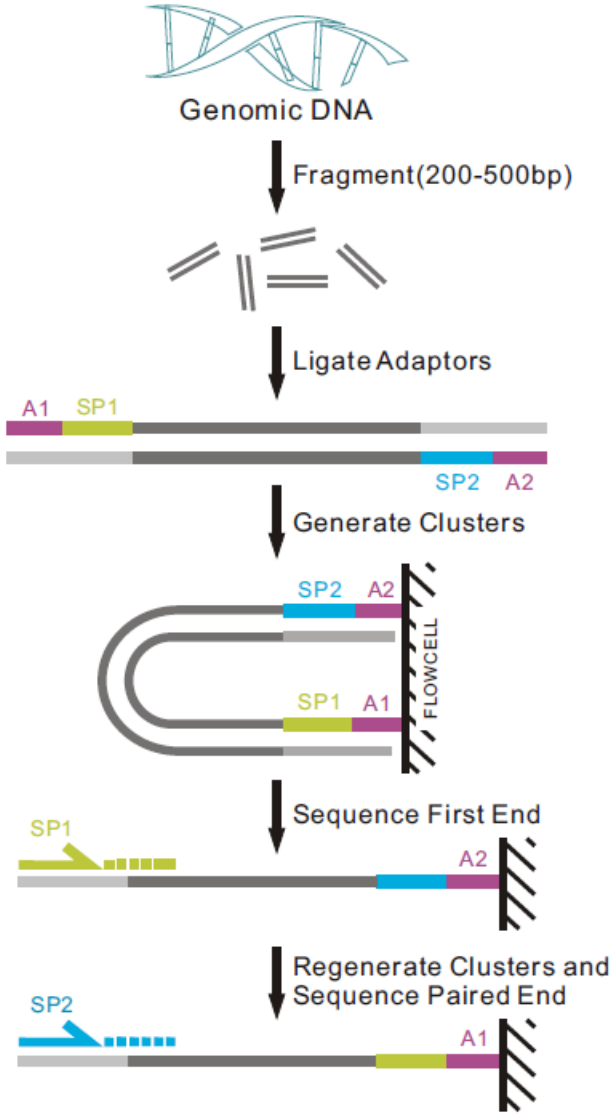
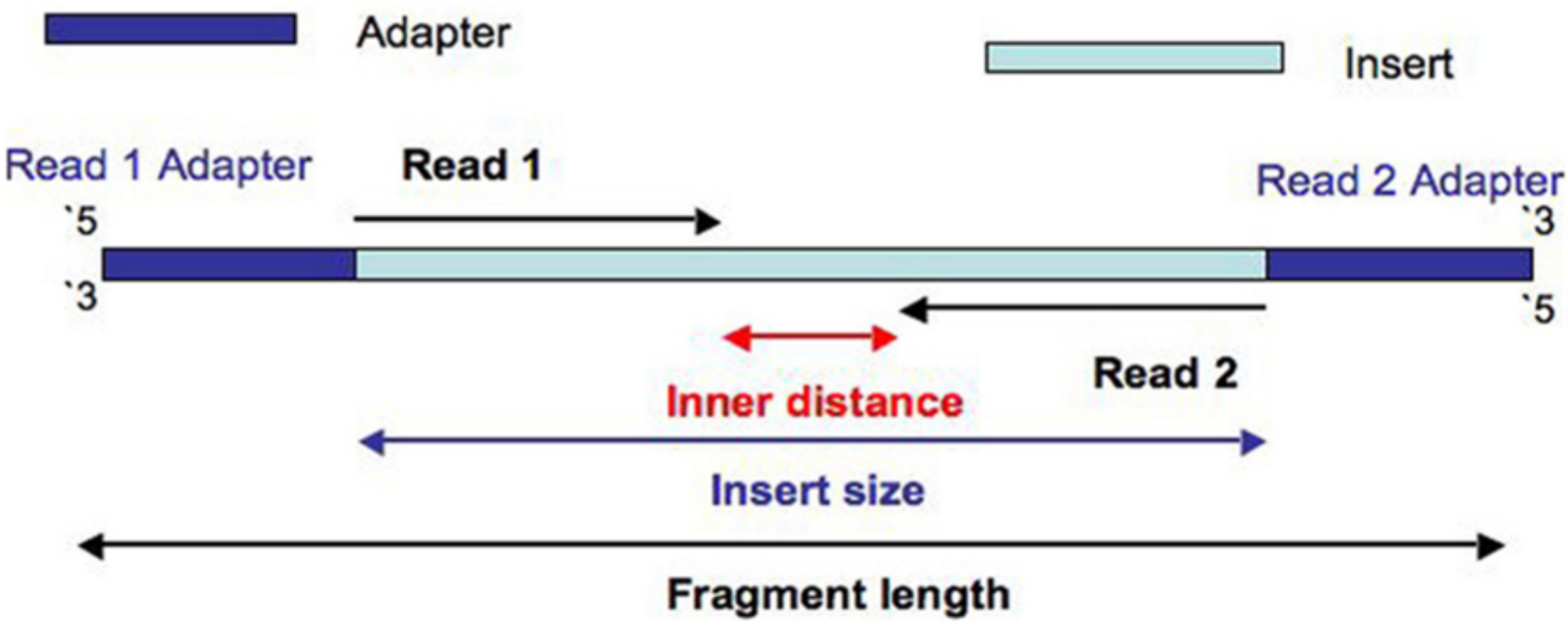


Figure 1-2-1 Pipeline of paired-end sequencing (www.illumina.com)


Adapter
Insert
Read 1 Adapter
Read 1
Read 2 Adapter
`5
`3
`3
`5
Read 2
Inner distance
Insert size
Fragment length

MAPPING TUTORIAL

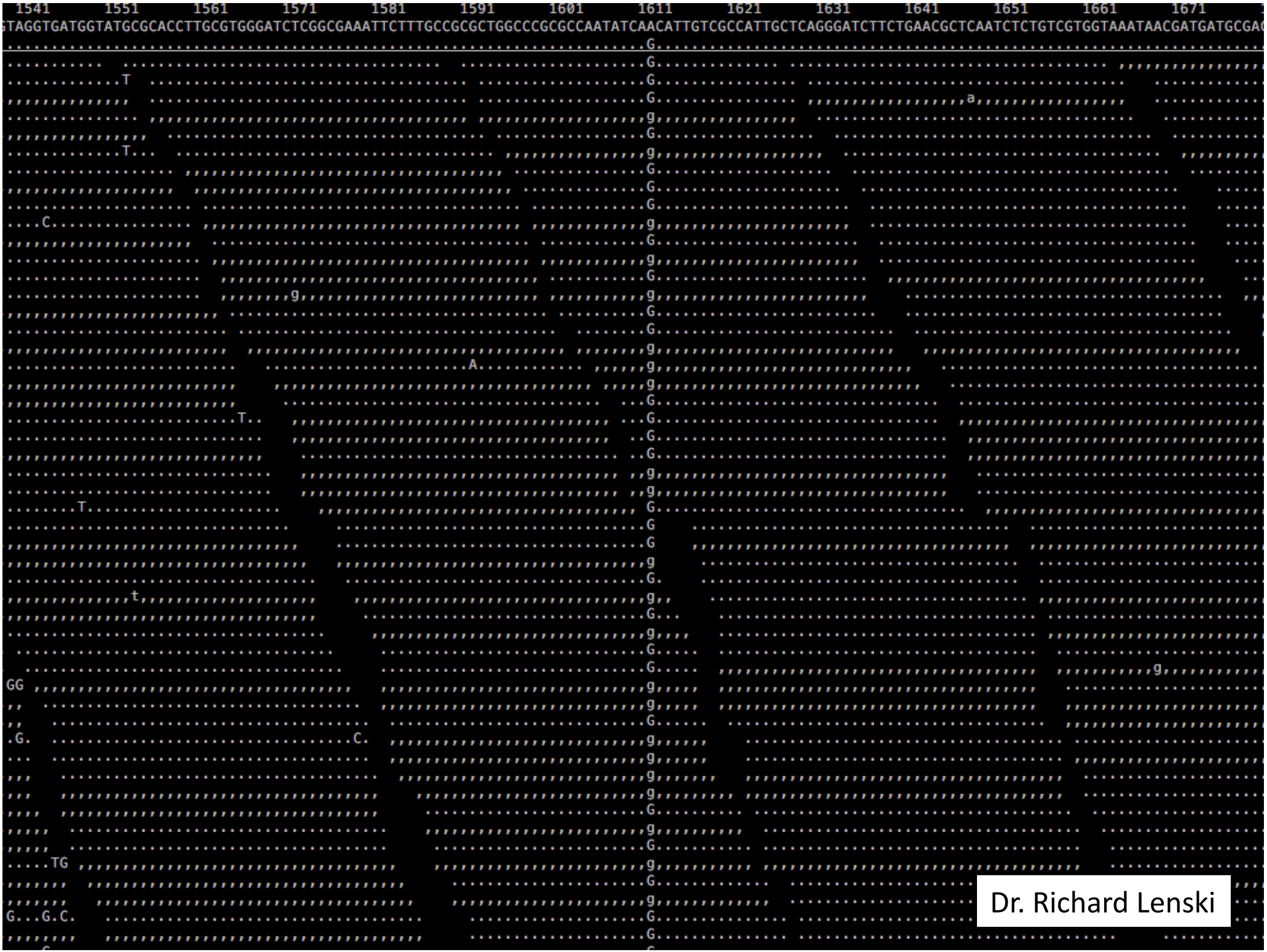
What is the SAM file?

- SAM stands for Sequence Alignment/Map format

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

What is the BAM file

- Binary Alignment/Map
- Smaller size than SAM file


Dr. Richard Lenski

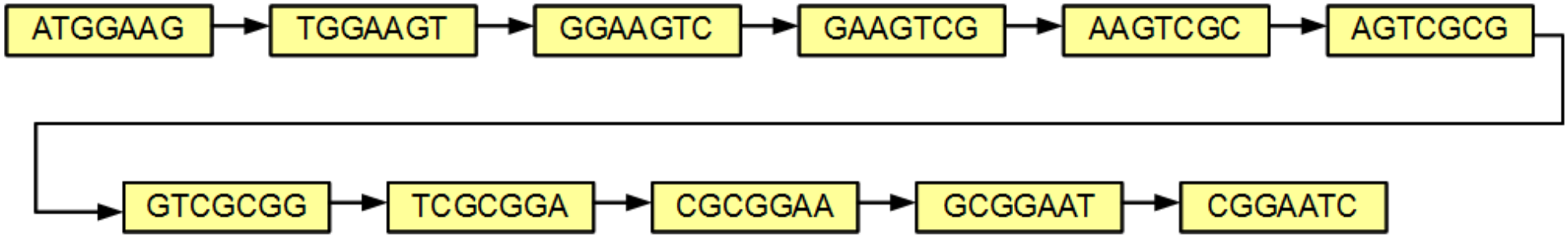
DE BRUIJN GRAPH

sequence **ATGGAAGTCGCGGAATC**

7mers

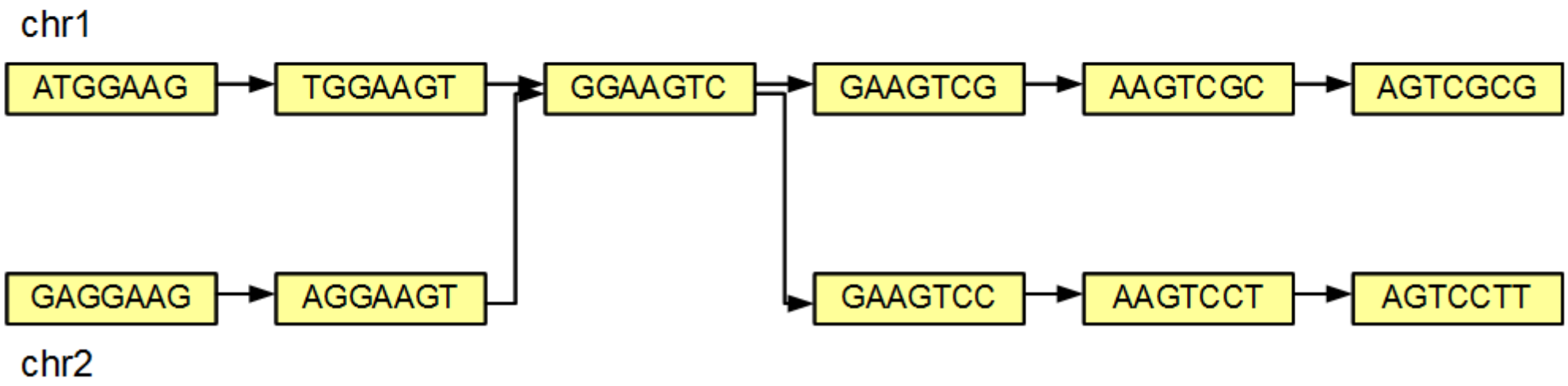
ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph



chr1 **ATGGAAGTCGCG**

chr2 **GAGGAAGTCCTT**



A

...ATTCCTGCAATAC...

...ATTCCTACAATAC...

ATT

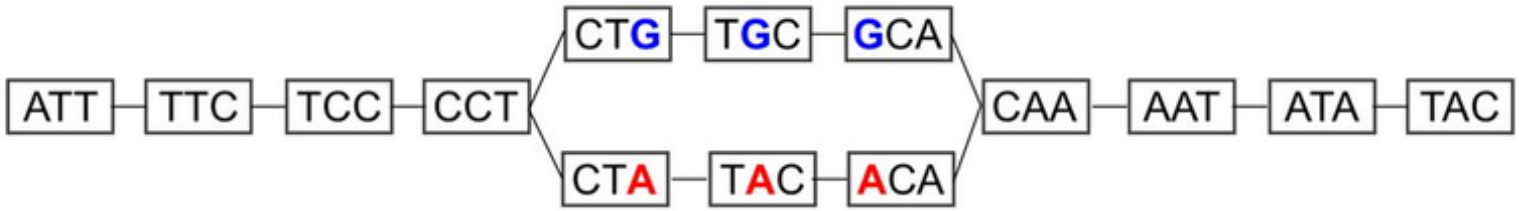
TTC

TCC

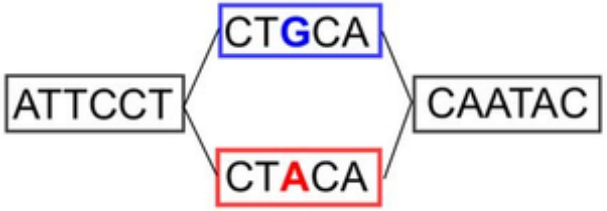
CCT

...

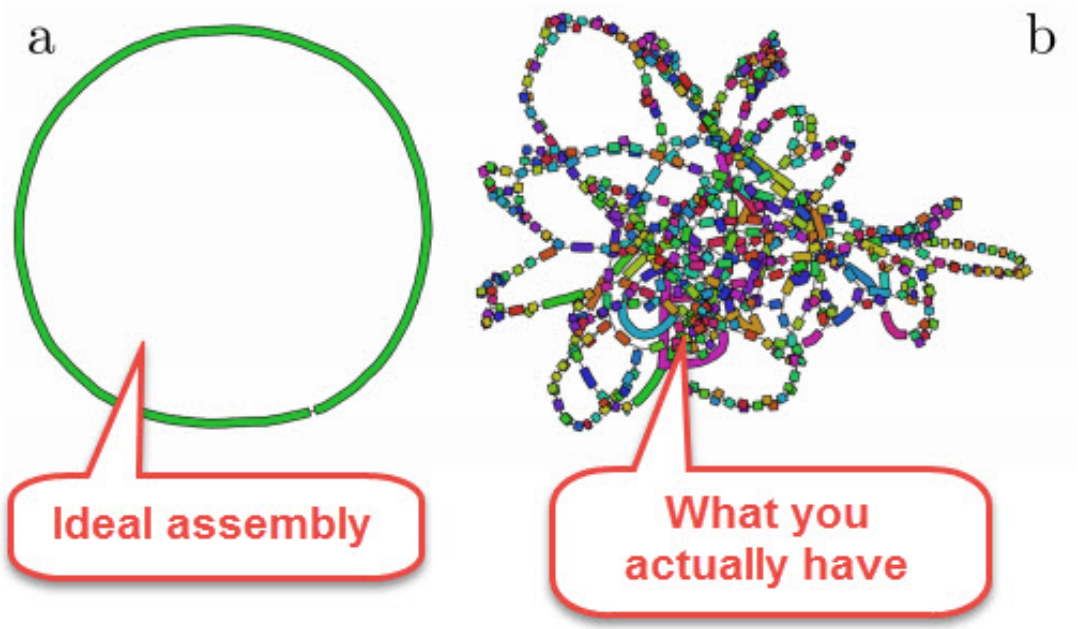
B



C



D


a
b
Ideal assembly
What you actually have